DOCUMENT RESUME

ED 047 319 AL 002 762

AUTHOR Marlin, Marjorie; Barron, Nancy
TITLE Methodology and Implications of Reconstruction and Automatic Processing of Natural Language of the Classroom.
INSTITUTION Missouri Univ., Columbia. Center for Research in Social Behavior.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Feb 71
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, N.Y., February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Classroom Communication, Computational Linguistics, *Computer Programs, *Data Processing, Grammar, *Research Methodology, Sociolinguistics, *Structural Analysis

ABSTRACT

This paper discusses in some detail the procedural areas of reconstruction and automatic processing used by the Classroom Interaction Project of the University of Missouri's Center for Research in Social Behavior in the analysis of classroom language. First discussed is the process of reconstruction, here defined as the "process of adding to messages what is otherwise not directly observable in the overt communicating behavior (i.e., grammatically and contextually implicit information), and of structuring this behavior into simplex sentences and lexical units." This is followed by a consideration of the role of the computer in processing the reconstructed data and a discussion of the computer programs used. The author believes that the language analysis system explained here has the advantage of universality, being applicable not only to classroom discourse but to any corpus of linguistic communication as well. See related documents AL 002 750-753. (FWB)

METHODOLOGY AND IMPLICATIONS OF RECONSTRUCTION AND AUTOMATIC PROCESSING OF NATURAL LANGUAGE OF THE CLASSROOM*

by

Marjorie Marlin
Nancy Barron

*The research leading to the results reported here have been supported by funds from the National Science Foundations, Grant No. GS3232 and by Institutional Research funds to the Center for Research in Social Behavior at the University of Missouri, Columbia, Missouri .

Our method of analyzing classroom language includes two procedural areas, reconstruction and automatic processing, which contain unexplored implications for language study. The propositions underlying the reconstruction system and the automatic processing algorithms will be discussed more or less separately although they are in fact interrelated.

The process of adding to messages what is otherwise not directly observable in the overt communicating behavior (i.e., grammatically and contextually implicit information),and of structuring this behavior into simplex sentences and lexical units, is referred to as reconstruction. In this definition, simplex sentences and lexical units are deemed the units of our analysis. Furthermore, our reconstruction system rests on the premises that implicit information is included in a communication, and for that communication to be properly analyzable this implicit information must be extricated from speech.

Roughly speaking, a simplex sentence is a clause, and the notion of reconstructing a text is analogous to that of parsing a sentence into its constituent clauses. Naturally occurring sentences may vary from a simple form in which we find a single noun and intransitive verb to a complex form in which multiple subjects objects, or verbs appear, with adjectival modifiers supplying additional meaning. In order to provide a simple format for analyzing meaning, it is useful to break down the natural sentence into a series of simple propositions that represent its meaning. The single verb propositions are simplex sentences. A lexical unit is a segment of

reconstructed text which is a nominal phrase, a verb phrase, a link, a nominalized sentence slot, or uncoded material. We have considered these the irreducible units of lexical analysis.

In effect, we have provided our own definitions for parts of speech. That is, we have asserted that any conversation can be represented by such lexical units as: a) nominals (including adjectives, articles, prepositions as well as nouns and pronouns), verbals (including adverbs as well as the main verb and its auxiliary structure), c) links (including most conjunctive particles and traditional connectors), d) nominalized sentence slots (which, as embedded sentences, fulfill the function of a phrase in a larger sentence unit), and e) uncoded materials, such as uh-uh, yes, well, good, etc.

In reconstruction , each lexical unit and each simplex sentence must be appropriately designated. This allows lexical units to be assigned to simplex sentences, and simplex sentences to be placed in order. To designate units, we adopted a scheme that made such identification precede the unit designated. Data were processed in sequential strings. One lexical unit was "ended" by the occurrence of the next identification number. This scheme demanded that all text "belong" to an identification number. That is, every word in the reconstructed text had to be assigned to its appropriate lexical unit number. Each naturally occurring sentence, each simplex sentence within it, and each lexical unit had to be continuous. Intervening structure within the simplex sentence was recorded by means of special conventions involving arbitrarily chosen punctuation marks.

Hierarchical claims about language underlie the process of separating natural sentences into lexical units. Some links are defined to represent a subordination of one simplex to another, as, for example, "unless" and "because." To consider clauses as optionally countable units implies an ordering, and asserts that the larger natural sentence is supraordinate to the subordinate simplex structures which underlie it. These assertions are built into the reconstruction procedures. Of course, the code makes judgments about complexity of sentence structure derived from a transformational grammatical theory.

Some implications of the lexical phrase and simplex sentence units relate to assumptions about the performance and competence of language users. It is assumed that members of a speech community share a common set of rules regarding appropriate, permissible utterances. It is further assumed that underlying these utterances are concepts that relate in some fashion to those employed by the grammarian to account for the competence of the speakers, such as simplex sentences and nominals.

We assume that speakers possess an organized set of concepts, the content of which is encoded and dispatched to the receivers, who in turn possess the organized concepts necessary to decode and understand the message content. Of course, people do not communicate perfectly; still, the greatest amount of communicative meaning should be captured if the system is constructed to represent messages as though they reflected these common conceptual organizations. If such organizations were not to be found, speech would be intuitive, idiomatic and infinitely variant.

Our processing of speech for analysis adhered to the following steps. The initial transcribing of the videotapes was done by secretaries who were untrained in linguistics. This process resulted in a rough transcription. A gross post-editing was performed, then a fine post-editing, by research assistants with linguistic training who specialized in ferreting out information; for instance, scarcely audible material, specific noverbal information, and features of the interaction such as the target of the utterance. These editors' interpretations provide control of veridicality. Trained undergraduates then reconstructed the text according to the system described in the manual referenced by Mr. Guyette. The reconstructed text then underwent a double reconstruction editing, first by a linguistically trained graduate student and then by the research coordinator (Dr. Barron). Ambiguities in interpretation were resolved . Judgments involved in assignment of identification numbers, by lexical unit, simplex sentence, and natural sentence classification, and by content of the lexical unit with regard to its form, its referent, etc., were refined.

Reconstructed texts were entered directly into the computer as a data source. Each lexical unit was reproduced as punched on a computer card, including the text of the unit itself, the natural sentence number, simplex sentence number, unit type (verb, link, etc.). This information was then reproduced by computer in three separate formats. First was a list format, in which lexical units appeared separately, one per line. Next was a straight text format, with simplex sentences appearing as text, one per line, including the punctuation conventions we had chosen to indicate such concept as implicitness or inserted referents (see Exhibits A&B). Third was an expanded text format, which

featured the addition of type identification before each lexical unit within the text. Information in this form was then used for the coding of lexical units and sentences.

The lexicon code concerned judgments about lexical units--words and phrases--such as their case function, pronominalization, gender, and implicitness or explicitness. The sentence code concerned judgments about language at the level of the simplex sentence, such as mode--question, command, assertion--and structural complexity type--adjoining, conjoining, and embedding mechanism. A check program was run on the reconstruction to find format and punctuation errors. Coding judgments were keypunched and verified. Finally the coding and reconstructed text were collated, referenced, and stored on magnetic tape for analysis.

Due to the sheer mass of the material, procedures were devised whenever possible to handle the process by computer. Utility programs were written to make corrections, to collate coding cards, and to ensure that information was reproduced on tape in precise columnar form. We now have on tape a total of 230 classroom minutes of reconstructed and coded sentences, comprising zpproximately 83,000 computer card images.

Some problem areas were painfully uncovered as we went along. For example, some arbitrary decisions with regard to coding proved to be less efficient than we had hoped. For instance, we would have included some convention to indicate the head noun or head verb of a lexical unit, had we known that such indication would later prove to be desirable. This information is now retrievable only through human judgment.

In the main, the restriction of back translatability of the reconstructed text to the fine-post-edited text has been adhered to. The exception--interrupted (noncontiguous) simplex sentences--will be accounted for in the next data processing by adding a new lexical unit designation. It is intended to computerize the back translation procedures and get a numerical measure of recoverability by back translation.

Difficulty has also been encountered in the sequential numbering of naturally occurring sentences. For example, if a sentence is accidentally "lost" in the assigning of identification numbers, then "found," all subsequent sentences must be renumbered. An expired classroom time designation might be a desirable substitution for sequential numbers, given additional equipment.

Now let us consider the computer programs available for processing reconstructed and coded data. First, programs were developed for preparing textual listings of data, as has been described above. Second, retrieval and classification of the data stored on tape has been done by programming the computer to output the specific information required. A set of programs has been implemented which produces frequency counts for specifically requested variables. We have a case count program which produces a table of frequencies of case use, cross-classified by "teacher" vs. "pupil" emitter. Another program gives us cross-classifications of items by gender, for animate cases only, cross-classified by implicitness vs. explicitness. Data from these programs have been analyzed and results are presented here by Dr. Barron.

One program produces a count of simplex sentences with respect to their complexity coding. These data have been subjected to

analysis for all segments of the sample, and the results are presented by Dr. Loflin.

Still another program produces lists of every reconstructed referent within a particular segment, together with a subsumed list of specific antecedents used for such referents. Frequencies of occurrence of each referent and their antecedents have been tabulated. Another program produces information as to the referent count, cross-classifying with respect to implicit vs. explicit occurrence and also with respect to pro-form vs. full-form occurrence. Results of analysis of these data are being presented by Mrs. Keyes and Mr. Guyette.

A dictionary and word count has also been computer-produced for each portion of the sample.

A third set of programs was devised to produce and collect information more detailed than simple frequency counts. For instance, it turned out to be difficult to compare sentences from different parts of a segment of the data in order to make judgments concerning similarity of meaning, inasumch as such sentences might be separated in time and space by a considerable amount of text. Hence a program was implemented to sort simplex sentences by person, number, and gender of nominal items within the simplex, and to output the sorted sentences. Human judgments about similarity of meaning were much facilitated by such sorting. Verbs have also been sorted in accordance with their co-occurrence with animate cases and gender, including a cross-classification of "self" and "other" reference. This information is presently under analysis by the authors.

Lists of case frames (patterns of cases within simplex sentences) occurring within the sample have been made, including a count of the

number of times a given case frame occurs, and also a list of the verbs occurring within such case frames.

Finally, an initial attempt to make a sequential analysis of data is underway. From information produced in the "dictionary" for each sample, a short list of content words of high frequency was selected. To enter the computer with such a list allowed us to derive a display of loadings for each sentence. A cyclical pattern of such loadings emerges, when one examines the occurrence of these high frequency words (see Exhibit C).

Work on such sequential analysis is now proceeding. As a first step in this analysis, we are in process of computing an entropy index for all the natural sentences. The entropy index was originally conceived to test an hypothesis about the structural characteristics of topic units. This hypothesis was derived from the postulate that subject matter was related to a set of structural (content-free) language characteristics: i.e., certain pro-form substitutions, implicitness, and lexical repetition. These indices share the common feature of a lack of new information, or redundancy. Thus they should load more or less additively on a common index of entropy. The hypothesis concerning topic variation stated that speech containing a new topic would be heralded by a burst of information, and then, sequentially, would be characterized by an increasing degree of entropy, or lack of new information, or repetitiveness. More succinctly, structural indices of language entropy were expected to vary in a cyclical fashion over time, and the cycles were expected to coincide with semantically-based judgments of topical units. Preliminarily, this relation seems in fact to exist; somewhere around 80% is the level of entropy which characterizes

a topic sequence. However, the entropy index has become fascinating in its own right, independent of topic. We expect to use it to document sociolinguistic characteristics of sequential speech patterns.

We have produced two versions of such an index. Originally we attempted to calculate the index by looking at the occurrence of chosen structural components of each sentence. These included 1) implicitness of the lexical item, 2) occurrence of pronoun substitution (that, which, what, etc., and personal pronouns such as he, it, etc.), 3) occurrence of referents which had appeared as explicit items earlier in the body of data, and 4) absence of "new words" in the lexical unit. The index was calculated as a ratio: out of the total number of lexical units which occurred in the natural sentence, what proportion were "redundant" because of any one or more of the four criteria above?

The original index was not entirely satisfactory for two reasons. First, there was a biasing toward the beginning of a selected body of data as an artifact of the initialization. In addition, the storage required of the computer on a long segment became prohibitive, since all words already occurring had to be stored. In addition, we desired to explore the possibility that better prediction of topic change was possible if the concept of additivity of the components was incorporated.

With the development of the "high frequency word" sentence loadings, it seemed that we might have here a substitute for the "new word" component of the entropy index. This has now been incorporated into the program which calculates the index. Because of this decision, calculation of the index has become a two-stage process: first the high-frequency words are selected from the dictionary for that segment of the data, then the computer checks each item for a redundancy load on each of the four criteria.

The program produces six indices for a given sentence: a ratio of items which are redundant on each count (of the four criteria), a proportion of item redundancy of any count, and finally an average redundancy including all loadings in an additive sense. Data from these indices are illustrated in Exhibit D.

It may be suggested that in the future some weighting of various structural components of natural sentences might be used for various purposes, such as the establishment of units of topic in classroom discourse. At present, equal weighting has been the only such scheme investigated, in the absence of any theoretical justification for a choice of weights on any other basis.

In principle, calculation of redundancy indices (or any other type of counting or sorting) might be done without computer processing. However, if we are to use the computer, it is necessary to make explicit the steps that are involved in such human judgments as counting and calculation, in order to translate such steps into instructions for the computer. It is the enormous mass of the data and the repetitive nature of many of the judgments involved which have dictated our extensive use of computer processing.

All programs mentioned here are written in PL-1 and have been implemented on the IBM 365-65 at the University of Missouri. It should be emphasized that these programs have been tailored specifically to the reconstruction and coding systems which we have devised--that is, they make use of the formatting and punctuation and labeling conventions used in our processing of the data. Most of these programs are fast-running and require computer capacity of 200K or less. Any could be rewritten

for a different reconstruction, coding, or computer system relatively easily.

We conceive of our language analysis system as applicable not only to classroom discourse, but to any corpus of linguistic communication. The universality of the system is one of its greatest advantages.

List Format

EXHIBIT A

```
                12          11          01          6          8          7          9          5          4          3
                                                                                                                      1
 37 02  'V'  'BEING SET UP'
 37 02  '2'  'IN FRANCE'
2X0037022 212010501                                 007
 37 02  '3'  '(BY THE REVOLUTIONARIES)'
2X0037025 1120101329                                007
 38 01  '1'  'THIS =37.XX/'
2X0038011 212020801                                 007
 38 01  'V'  'LASTED'
 38 01  '2'  'FOR A SHORT TIME'
2X0038012 220110701                                 007
 39 01  '1'  'THE JEFFERSONIANS'
2X0039011 2120102329                                007
 39 01  'V'  'WERE GREATLY PLEASED'
 39 01  '2'  'WITH THIS =37.XX/'
2X0039012 212021501                                 007
 39 02  'L1' 'BECAUSE'
 39 02  '1'  'THEY =JEFFERSONIAN;/'
2X0039021 2120201329                                007
 39 02  'V'  'LOOKED'
 39 02  '2'  'TO FRANCE'
2X0039022 2120103323                                007
 39 03  'L1' 'AS (THOUGH)'
 39 03  '1'  'IT =FRANCE/'
2X0039031 2120202323                                007
 39 03  'V'  '(WERE)'
 39 03  '2'  'THE GREAT CHAMPION'
2X0039032 2109102323                                007
 39 03  '3'  'OF DEMOCRACY'
2X0039033 222011501                                 007
```

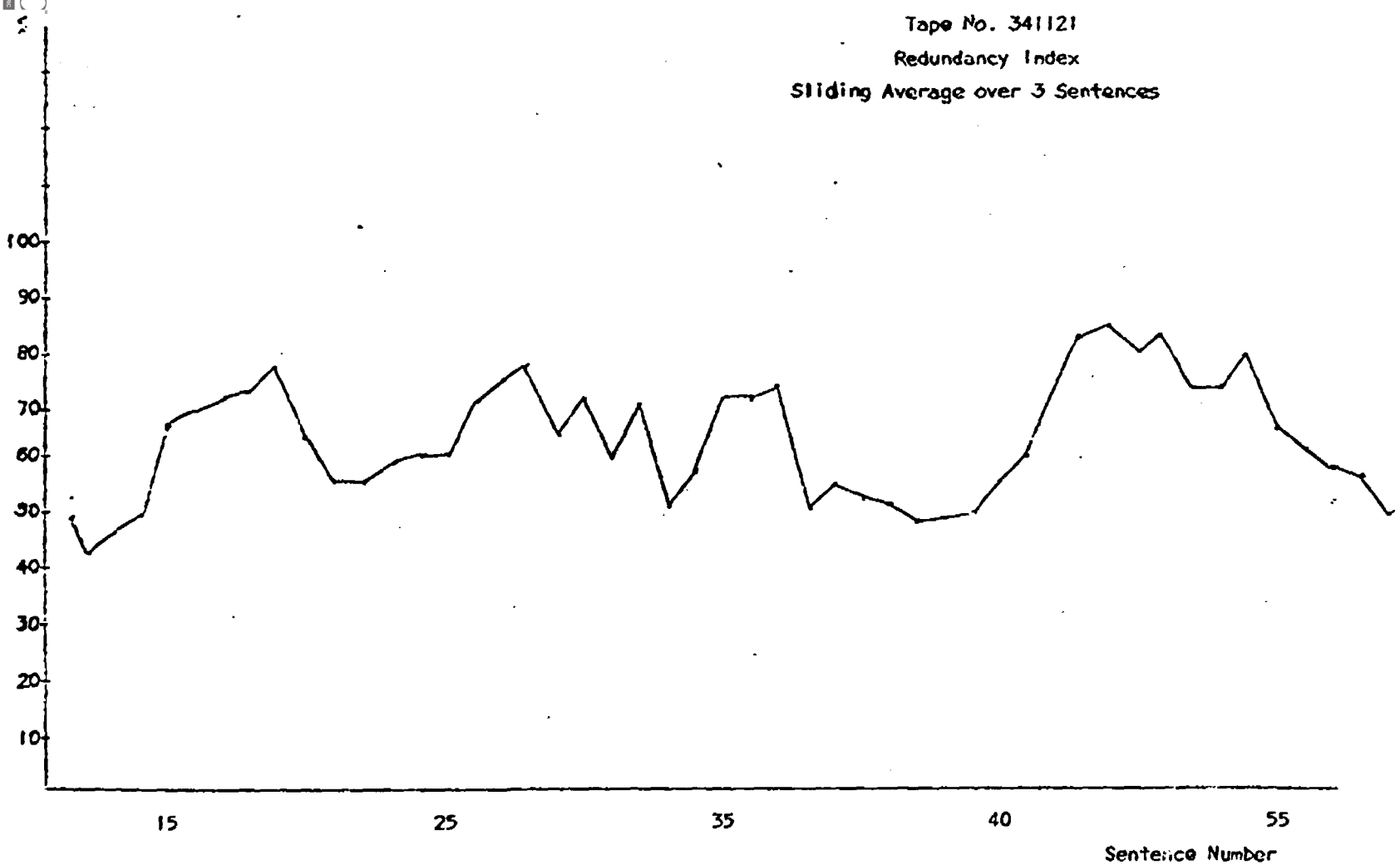
```
 1  AND YOU =?/ HAVE//SUPPOSEDLY NOW *02
 2  THIS DEMOCRATIC GOVERNMENT BEING SET UP IN FRANCE (BY THE REVOLUTIONARIES)
 1  THIS =37.XX/ LASTED FOR A SHORT TIME
 1  THE JEFFERSONIANS WERE GREATLY PLEASED WITH THIS =37.XX/
 2  BECAUSE THEY =JEFFERSONIANS/ LOOKED TO FRANCE
 3  AS (THOUGH) IT =FRANCE/ (WERE) THE GREAT CHAMPION OF DEMOCRACY
 1  HOWEVER THIS =39.XX/ WAS SOON DONE AWAY WITH (BY SOMEONE =?/)
 2  BECAUSE BY (THE YEAR) SEVENTEEN NINETYTWO # # (OR)
 3  (YOU =PUPILS/) REMEMBER *06
 4  BECAUSE OF THE ARREST OF THE FRENCH NOBILITY (BY THE REVOLUTIONARIES) AND
 5  (BECAUSE) ALREADY THE BEHEADING OF MARIE ANTIONETTE (BY THE REVOLUTIONARIES)
 6  YOU =?/ HAVE *07,08,09
 7  THE HAPSBURGS OF AUSTRIA (DECLARE) (WAR) (ON FRANCE) (AND)
 8  (THE HAPSBURGS) (OF) HUNGARY (DECLARE) (WAR) (ON FRANCE) AND
 9  ALSO//DECLARE SPAIN WAR ON FRANCE
10  WHICH =SPAIN/ IS ALSO CONTROLLED BY THE HAPSBURGS
 1  AND THEREFORE NOW YOU =?/ BEGIN TO HAVE *02,03
 2  A EUROPEAN WAR TAKING PLACE =OCCURRING/ (AND)
 3  (A EUROPEAN WAR) BEGINNING IN (THE YEAR) SEVENTEEN NINETYTWO
 1  (YOU =PUPILS/) REMEMBER *02,03
 2  THAT LOUIS THE SIXTEENTH IS BEHEADED (BY THE REVOLUTIONARIES)
    IN (THE YEAR SEVENTEEN) NINETYTHREE AND
 3  THIS =42.02/ BRINGS ENGLAND INTO THE PICTURE
 4  BECAUSE SHE =ENGLAND/ IS NOT ONLY IN CONFLICT WITH FRANCE
 5  BECAUSE OF TRADE (BY ENGLAND) (WITH FRANCE) AND
 6  (BECAUSE OF) POSITION AS =IS/ THE MAJOR LAND POWER IN EUROPE BUT
 7  FRANCE'S (HAS) (POSITION)
 8  ALSO//LOOKS AT ENGLAND THIS BEHEADING =42.02/
 9  AS (THOUGH) (IT =42.02/) (WERE) A THREAT TO =THREATENED/ THE MONARCHIAL SYSTEM
10  BECAUSE (YOU =PUPILS/) REMEMBER *12
11  THE ENGLISH HAVE//TOO A KING AND
12  THEREFORE YOU =?/ HAVE NOW A EUROPEAN WAR
13  WITH =IN WHICH WAR/ AUSTRIA (VERSUS =WAS FIGHTING AGAINST/) (FRANCE) (AND)
14  (WITH =IN WHICH WAR/) HUNGARY (VERSUS =WAS FIGHTING AGAINST/) (FRANCE) AND
```

1 2 3 4 5 6 7 8 9 10

EXHIBIT C

FOLIO	Word Classes									Total Per Sentence	Totals for 3 Sentences
126											
127		1			2		1	1		5	5
128		2			3	4	2	2		8	13
129		2			5	4	3	3		4	17
130		2			8	4	2	2		6	18
131		1			7		1	1			10
132		2			5					1	7
133		1									1
134		2								1	2
135		1									1
136		1									1
137											
138		1								1	1
139		1									1
140		2								1	2
141	1	1			2					3	4
142	1	3	2	2	2					6	10
143	1	3	4	2	2					3	12
144		5	6	4						6	15
145		3	4	2							9
146		3	4	2		1				4	10
147		2	4			1				3	7
148		2	5			1				1	8
149	2	1	3							2	6
150	2		2							1	4
151	2	1	1	2		1				4	7
152		2	1	4		2				4	9
153		3		6		3				4	12
154		2		4		2					8
155		1		2		1					4
156	2	2	4	8	2	1				17	17
157	2	4	8	6	2	1				6	23
158	2	5	8	12	2	1				7	30
159		4	4	6				1		2	15
160		2		6				1			8
161		2						1		1	3
162		1			2	2				4	5
163	2	2			2	2	1	3	3	10	15
164	2	3			2	2	4	3	3	5	19
165	2	4				5	6	6	6	14	29
166		3				5	5	3	3		19
167		1				5	2	3	3		14
168					2					2	2
169	1	1			2	1				3	5
170	3	1			2	1				2	7
171	1 3	1			1	1	9			11	16
172	1 4				5	2	9			8	21
173	3 3				6	3	9			5	24
174	2 3				6	9		1	1	9	22
175	2 1	1			2	7		1	1	1	1
176	8	2			1	6		1	1	9	19
177	9	2								1	11
178	9	1						1		1	11
179	1							1			2
180								1			1
181	5	6			7		8		9	10	

EXHIBIT D


Tape No. 341121
Redundancy Index
Sliding Average over 3 Sentences
100
90
80
70
60
50
40
30
20
10
15
25
35
40
55
Sentence Number